A multi-data set comparison of the vertical structure of temperature variability and change over the Arctic during the past 100 years

Brönnimann, S ; Grant, A N ; Compo, G P ; Ewen, T ; Griesser, T ; Fischer, A M ; Schraner, M ; Stickler, A

Abstract: We compare the daily, interannual, and decadal variability and trends in the thermal structure of the Arctic troposphere using eight observation-based, vertically resolved data sets, four of which have data prior to 1948. Comparisons on the daily scale between historical reanalysis data and historical upper-air observations were performed for Svalbard for the cold winters 1911/1912 and 1988/1989, the warm winters 1944/1945 and 2005/2006, and the International Geophysical Year 1957/1958. Excellent agreement is found at mid-tropospheric levels. Near the ground and at the tropopause level, however, systematic differences are identified. On the interannual time scale, the correlations between all data sets are high, but there are systematic biases in terms of absolute values as well as discrepancies in the magnitude of the variability. The causes of these differences are discussed. While none of the data sets individually may be suitable for trend analysis, consistent features can be identified from analyzing all data sets together. To illustrate this, we examine trends and 20-year averages for those regions and seasons that exhibit large sea-ice changes and have enough data for comparison. In the summertime Pacific Arctic and the autumn eastern Canadian Arctic, the lower tropospheric temperature anomalies for the recent two decades are higher than in any previous 20-year period. In contrast, mid-tropospheric temperatures of the European Arctic in the wintertime of the 1920s and 1930s may have reached values as high as those of the late 20th and early 21st centuries.

DOI: https://doi.org/10.1007/s00382-012-1291-6

A multi-data set comparison of the vertical structure of temperature variability and change over the Arctic during the past 100 years

Stefan Brönnimann[1,2], Andrea N. Grant[1], Gilbert P. Compo[3,4], Tracy Ewen[5], Thomas Griesser[1], Andreas M. Fischer[6], Martin Schraner[1,7], Alexander Stickler[1,2]

[1] *Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland*

[2] *Oeschger Centre and Institute of Geography, University of Bern, Switzerland*

[3] *Climate Diagnostics Center, CIRES, University of Colorado, Boulder, USA*

[4] *Physical Sciences Division, Earth System Research Laboratory, NOAA, Boulder, USA*

[5] *Department of Geography, University of Zurich, Switzerland*

[6] *Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland*

[7] *Swiss National Supercomputing Centre CSCS, Manno, Switzerland*

Corresponding author:

Stefan Brönnimann

Oeschger Centre for Climate Change Research and Institute of Geography

University of Bern

Hallerstr. 12

CH-3012 Bern, Switzerland

e-mail: broennimann@env.ethz.ch

Phone: ++41 031 631 88 85

Fax: ++41 031 631 85 11

Abstract

We compare the variability and trends in the thermal structure of the Arctic troposphere in eight observation-based, vertically resolved data sets, four of which have data prior to 1948. Comparisons on the daily scale between historical reanalysis data and historical upper-air observations were performed for Svalbard for the cold winters 1911/1912 and 1988/89 and the warm winters 1944/1945 and 2005/2006. Excellent agreement is found at mid-tropospheric levels. Near the ground, however, systematic differences are identified. On the interannual time scale, the correlations between all data sets are high, but there are systematic biases in terms of absolute values as well as discrepancies in the magnitude of the variability. The causes of these differences are discussed. With respect to overlapping 20-yr averages and trends in the vertical structure, the data sets also agree well, although differences are more pronounced than for the interannual scale. While none of the data sets individually may be suitable for trend analysis, consistent features can be identified from analyzing all data sets together. To illustrate this, we examine trends and 20-yr averages for those regions and seasons that exhibit large sea-ice changes and have enough data for comparison. In the summertime Pacific Arctic and the autumn eastern Canadian Arctic, the range of estimates for lower tropospheric temperature anomalies for the recent two decades does not overlap with any estimate for any previous 20-yr period. In contrast, mid-tropospheric temperatures of the European Arctic in the wintertime of the 1920s and 1930s may have reached values as high as those of the late 20th and early 21st centuries.

1. Introduction

Recently developed 4D data sets and reanalysis products spanning the 20th century offer the promise of new insight into the dynamics of climate variations in the past. A prominent example is the early 20th warming (ETCW, see also Brönnimann 2009); a period with pronounced warming in several regions, including the North Atlantic, with a particularly large amplitude in the Arctic (e.g., Polyakov et al. 2003, Bengtsson et al. 2004, Overland et al. 2004, Johannessen et al. 2004, Wang et al. 2007, Kauker et al. 2008, Wood and Overland 2010, Wood et al. 2010). A study of the vertical structure of the warming in the Arctic troposphere in these new datasets might give indications as to the relative roles of atmospheric heat transport and processes operating near the ground (see Graversen et al. 2008, Serreze et al. 2009, Screen and Simmonds, 2010 for corresponding studies on the ongoing warming). Existing datasets for later periods have problems in this respect (Bromwich and Wang 2005, Thorne 2008, Grant et al. 2008, Bitz and Fu 2008, Screen and Simmonds 2011). The data quality and suitability of the new, long data sets that cover the ETCW have not been assessed.

The main goal of this study is to assess and intercompare the newly-available global, 3-dimensional observation-based temperature data sets with respect to their representation of Arctic tropospheric temperature during the twentieth century. In order of period covered, these are: The Twentieth Century Reanalysis (20CR; 1871-2008), two statistical reconstructions (REC1; 1880-1957 and REC2; spatially incomplete, with Arctic data from 1923-1957), and upper-air observations (CHUAN, spatially incomplete, with Arctic data from 1930-2006). These data sets are supplemented with some widely used reanalysis data sets, i.e., NCEP/NCAR Reanalysis (NNR, 1948-2009), ERA-40 reanalysis (1957-2002), JRA25 (1979-2009), and ERA-Interim (1989-2009, see also Fig. 1 and Table 1).

Comparisons are performed for different Arctic regions and seasons, but for three reasons special emphasis is devoted to the European Arctic, particularly Svalbard. First, the European Arctic is believed to be a critical region for our understanding of Arctic climate processes (e.g., Bengtsson et al. 2004, Pethoukov and Semenov 2010). Second, this region of the Arctic exhibits particularly high temperature variability on synoptic to interannual scales (see, e.g., Grant et al. 2009b). The considered atmospheric data sets should be capable of capturing this variability. Finally, historical upper-air observations are available for Svalbard. Though sparse and heterogeneous, they nonetheless form one of the longest Arctic records which exist for such analysis.

By assessing and intercomparing the new datasets, several findings are made possible that would be only suggestive if any single dataset were used. In some Arctic regions, the recent warming is commensurate with warm anomalies seen during the ETCW. In other regions, however, the most recent 20 year period of lower tropospheric warming is extraordinary, both in its magnitude and in its lapse rate, compared to any prior period of the 20th century.

The remainder of the paper leading to these findings is organized as follows. Section 2 gives a description of the data used. The concept and methods are outlined in Section 3. In Section 4 we show the results of the comparison and discuss prominent features of warm periods and trends in the Arctic troposphere. Conclusions are drawn in Section 5.

2. Data

Eight different upper-air datasets are included in this assessment (Table 1). As described below, these are: observed data (a), statistically reconstructed data (b), and reanalysis data (c) that have commonalities and differences in their generation that should be kept in mind when interpreting the results.

a. Observations

As a reference for our comparisons, we use observational datasets, keeping in mind that measurements and averages based on them contain errors. To represent the near-surface air temperature, we use the gridbox anomaly dataset of CRUTEM3v (Brohan et al. 2006). We also use temperature station data from Svalbard from the NORDKLIM project (Tuomenvirta et al. 2001) updated after 2001 using NASA/GISS data (Hansen et al. 1999).

Above the Earth's surface we use the temperature observations from a combination of radiosonde, kite, and aircraft-based measurements contained in the Comprehensive Historical Upper Air Network (CHUAN, Stickler et al. 2010, Grant et al. 2009a, Brönnimann 2003). An overview of the stations north of 60 °N is given in Fig. 2. Apart from some scattered data, the earliest records start in the 1930s, mainly from the former Soviet Union and from Scandinavia. Upper air records from the western hemisphere start mostly later, in the 1940s or 1950s.

The upper air data were quality assessed following Grant et al. (2009a). Corrections were applied up to the end of 1957. The series in CHUAN were supplemented for the period from

1958 to present using data from the Integrated Global Radiosonde Archive (IGRA, Durre et al. 2006) with RAOBCORE version 1.4 corrections (Haimberger 2007). With very few exceptions (see Stickler et al. 2010, for details) no new stations were added from 1958 on.

For Svalbard, in addition to the records found in IGRA (e.g., Ny Ålesund, see Fig. 2) and CHUAN (e.g., Barentsburg), we digitised further historical upper-air data from tethered balloons and kites from Advents Bay and Ebeltofthamna, 1911-1913, as well as radiosonde data from Nordaustlandet from 1944-1945, respectively, both performed by German observers.

The Advents Bay/Ebeltofthamna data were originally published by Rempp and Wagner (1916), Wegener (1916) and Wegener and Robitzsch (1916a,b). The balloons often did not reach very high altitudes, however, during the 22 months of measurements, 80 profiles reached an altitude of 1500 m asl (approximately 850 hPa).

The data from 1944/1945 are from the German war operation "Haudegen" led by Wilhelm Dege (Selinger 2001). In total 132 radiosonde ascents were performed between November 1944 and June 1945. Pilot balloon observations were also made (until Sep. 1945, when the station was finally uncovered, making this the last German unit to surrender), but not used in this project. We used radiosonde temperature data on standard pressure levels as given in Dege (1960). The source does not mention whether radiation and lag error corrections have been applied. Since the data were published in 1960, we assume that these errors were in fact corrected. We also tested the possible bias from using uncorrected data (following Brönnimann 2003) and found that it would lie between -1 °C and +0.3 °C (depending on the ascent and level; the average over all ascents and levels considered here is -0.33 °C). For the winter period (a focus of this paper), when the radiation errors are small, the bias is even more reduced.

Periods of available upper-air data series from Svalbard are shown in Fig. 1, together with 850 hPa temperature in winter as an example. Data are available from many sites, but in the first decades they are very spotty (see Section 3b for the calculation of seasonal averages).

Note that both the tethered balloon data and the radiosonde data have various sources of uncertainties. These might be particularly large in the harsh Arctic environment. Unfortunately, we have no estimation of the error for these specific Arctic sites. A recent paper (Brönnimann et al. 2011b), estimates the error for early ship-based upper air data measured with kites and radiosondes. Here, we assume that random errors are of a similar magnitude of about 1 °C, in addition to the biases such as those mentioned above.

b. Reconstructions

Temperature fields for the period 1880-1957 are taken from a statistical reconstruction based on a principal component regression (Griesser et al. 2010). The predictors are historical surface data from station observations (temperature), gridded sea-level pressure (SLP), as well as upper-air data (temperature, geopotential height (GPH) or pressure, and winds) after 1918. The predictands used were hemispheric GPH and temperature fields at six levels (850, 700, 500, 300, 200, 100 hPa) from ERA-40 reanalysis (Uppala et al. 2005). The statistical models are calibrated in the period 1958-2001 and optimized using split sample validations within that period. This reconstruction is termed REC1. As an example, Fig. 1 shows 850 hPa winter temperature from REC1 interpolated to Svalbard.

A second reconstruction, REC2, avoids the strong limitations of constraining stationary patterns (large-scale empirical orthogonal functions) and thus stationary teleconnections (see Brönnimann et al. 2011a, for details). The approach of REC2 is similar to REC1 except that it is performed grid column by grid column (rather than with hemispheric fields) using only predictors in the "cone of influence" of that grid column (radius of 1200-1500 km depending on the variable and level, thus avoiding calibration by means of a possible negatively correlated series). This alleviates the need for stationary patterns, at the expense of a sparse data set. REC2 provides temperature, GPH, u and v winds at six levels (850, 700, 500, 300, 200, 100 hPa). It covers the period 1918-1957, but in the Arctic data start only in the 1920s (see Fig. 1). After 1957 the data set is continued using the predictor network from 1957 (denoted REC2-cal., see Fig. 1). Although that part of the data set is still based on observations, it is closer to ERA-40 reanalysis because it covers the calibration period and because gaps in the predictors after 1957 are filled with data extracted from ERA-40 (see Brönnimann et al. 2011a for details).

Both reconstructions use upper-air data from CHUAN and hence are not independent from CHUAN. However, a large amount of the Arctic upper-air data in CHUAN did not enter the reconstruction because monthly mean values could not be calculated on a station-by-station basis (a requirement for REC1 and REC2), whereas the method used in this paper to derive seasonal-regional averages from CHUAN makes use of all data. Also, both reconstructions give some information on the reconstruction skill.

c. Reanalysis data sets

Currently only one reanalysis data sets covers the ETCW in the Arctic. The Twentieth Century Reanalysis version 2 (20CR) is a global 3-dimensional atmospheric dataset that reaches back to 1871 (Compo et al. 2011). It is based on an assimilation of surface observations of synoptic pressure. HadISST (Rayner et al. 2003) monthly sea surface temperature (SST) and sea ice distributions are prescribed as boundary conditions. Time-varying radiative forcings of CO_2, volcanic aerosols, and solar output are also prescribed. Assimilation is performed using an Ensemble Kalman filter with first guess fields generated by a 2008 experimental version of the US National Centers for Environmental Prediction Global Forecast System atmosphere/land model (NCEP/GFS) at a spatial resolution of T62, with 56 ensemble members. Because it is an ensemble system, 20CR not only provides 6-hourly global analyses (ensemble mean) but also their uncertainty (the ensemble standard deviation). Details and validation results are given in Compo et al. (2011).

In order to better assess biases and differences, we compare the other data sets with four widely used reanalysis data sets (termed "conventional reanalyses" in the following): NCEP/NCAR (NNR hereafter) from 1948 to 2009 (Kistler et al., 2001), ERA-40 from 1958 to 2002 (Uppala et al., 2005), JRA-25 from 1979 to 2007 (Onogi et al., 2007), and ERA-Interim from 1989 to 2007 (Dee et al. 2011). Note that these data sets, too, have errors. Errors and inconsistencies in the assimilation system or in the data assimilated can lead to inhomogeneities and errors. Errors relevant for the Arctic include a warm bias in NNR over the former Soviet Union in 1948–1957 due to uncorrected radiation errors in the radiosonde data (Grant et al. 2009a). In the case of ERA-40, problems with satellite radiance assimilation over the ice-covered Arctic Ocean are documented (Bromwich and Wang 2005, Uppala et al. 2005), which can lead to spurious trends (e.g., Thorne 2008, Grant et al. 2008).

Conventional reanalyses use surface as well as upper-air input and hence are not fully independent from any other data sets during the period of overlap. 20CR, however, is completely independent from CHUAN. With REC1 and REC2 it shares some SLP input.

3. Analysis procedure

The eight data sets are compared with respect to their representation of the variability of temperature at different levels in the atmosphere. We analyse correlations to measure differences in variability on different time scales, averages to measure differences in the mean,

and trends to measure differences in the tendencies. We also analyse the consistency of observed and expected differences between datasets. Finally, we address the vertical structure of warm periods and warming trends across the eight data sets. Because upper-air observations form the reference for all comparisons but are themselves very sparse in the first half of the twentieth century, the comparison methods are strongly guided by the availability of observations.

a. Day-to-day variability in Svalbard

The agreement of data sets on the day-to-day scale can only be analysed for CHUAN and 20CR. We show results for the case of Svalbard, where CHUAN data also allow a mutual comparison of neighbouring observational data records. To facilitate comparison we subtracted a common climatology from each data set. We used NNR data for this purpose, namely a climatology of daily mean values as a function of the day of year that is given and recommended on the website of Physical Sciences Division of NOAA's Earth System Research Laboratory and refers to the period 1968-1996 (note that for the comparisons of the interannual variability, where more data sets than NNR are involved, we use 1961-1990 as a reference). These data also were subsampled and interpolated to the location and time of the ascents.

We also investigated the consistency of the data sets given their errors, as in Brönnimann et al. (2011b). We assumed that the standard deviation of the differences between upper-air observations and 20CR, σ_{diff}, can be estimated by the square root of the sum of three error terms (represented by their variances), i.e., the error of 20CR (we use the ensemble spread here), the error of the observations (we assume 1 °C following Brönnimann et al. 2011b), and the error of representativeness which is related to the interpolation in space and time (we assume 1.96 °C following Brönnimann et al. 2011b, for all cases):

$$\hat{\sigma}_{diff} = \sqrt{\sigma_{20CR}^2 + \sigma_{obs}^2 + \sigma_{rep}^2}$$

If 95% of the differences between CHUAN and 20CR are within ±2 $\hat{\sigma}_{\mathrm{diff}}$, they are consistent with the specified errors. Note, however, that this does not account for mean biases.

b. Interannual variability

Interannual variability was addressed for different regions of the Arctic and different seasons using monthly and seasonal-regional averages. Due to the sparseness of upper-air observations, which are used as a reference, the procedure of forming these averages was determined mainly by data availability. Not only is the number of observations small prior to the 1950s, they are also very heterogeneous (short records from many different sites, each with many gaps), as can be seen in Fig. 1 for the case of Svalbard.

Therefore, to use all observations as in Grant et al. (2009b), the following procedure was employed. The region poleward of 60 °N was divided into 54 equal area grid cells (Fig. 2), and time was subdivided into weeks. Both the grid cell size of approximately 800 km x 800 km and the seven-day blocks were chosen as representative of the intraseasonal large scale in order to maximize the information contained in the spatially and temporally sparse measurements. Anomalies of individual soundings were calculated relative to a 1961-1990 monthly climatology from NNR for each location and then averaged within the equal area grid cells and seven-day blocks. The mean values per grid cell and week were then aggregated into sectors and seasons.

The four seasons were defined as the periods of 1 December to 1 March (winter), 1 March to 31 May (spring), 1 June to 31 August (summer) and 1 September to 1 December (autumn). The overlaps (1 March, 1 December) are necessary for obtaining an integer number of weeks (thirteen) for averaging.

Despite making best use of all available observations, many of the grid cells still have too few observations and therefore existing regionalizations of the Arctic such as those by Treshnikov (see Przybylak, 2007) cannot be used. Rather, we defined regions as sets of 4-6 neighbouring grid cells with good in-situ data coverage. Seven regions with reasonable coverage can be identified. For brevity's sake we show figures only for four sectors (Fig. 2), each for one season, namely (1) the European Arctic in winter, (2) the Western Siberian Arctic in spring, (3) the Pacific Arctic in summer and (4) the eastern Canadian Arctic in autumn (see Fig. 2 for definition). These combinations capture different characteristics of Arctic climate. Moreover, combinations (1), (3), and (4) correspond to regions and seasons with a large variability in sea ice. Regions (1) and (2) correspond very roughly to western and eastern parts within Treshnikov's Atlantic Arctic region (but all regions reach further south than Treshnikov's), (3)

and (4) can best be compared with his Pacific Arctic and Canadian Arctic regions. Note that the Arctic Ocean is underrepresented and land areas are overrepresented in this selection.

Seasonal-regional means were then calculated from the grid cell averages if 50% of the grid cells in a region and 7 out of 13 weeks in the season had data. For the gridded data sets we simply averaged the region for the sectors as shown in Fig. 2 and used climatological seasons rather than to subsample all data sets to the exact times and locations of the observations (as it was done for the Svalbard station data in the previous section). This facilitates clearer interpretation of trends in the gridded datasets (whereas the sub-sampling would "transfer" uncertainties in the observational data, e.g., from changes in locations, to other data sets). However, with respect to the assessment of errors, it should be kept in mind that the differences between CHUAN and other products also contain the sampling error in addition to the errors addressed in the previous section.

We show seven levels, namely 1000, 850, 700, 500, 400, 300, and 200 hPa. Surface air temperature (from CRUTEM3v, Brohan et al. 2006) is shown rather than 1000 hPa from CHUAN, which is often extrapolated or not reported (CRUTEM3v data are also shown together with REC1 and REC2 which do not have the 1000 hPa level). Due to irregular reporting, the 925 and 600 hPa levels were omitted in the CHUAN averages. Similarly to the day-to-day variability, we analyze the regional-seasonal averages in the form of anomalies. For this purpose, the mean annual cycle from the years 1961-1990 was subtracted. All analyses were performed using both NNR and ERA-40 as a common reference as well as using each data set as a self-reference (only for long data sets). Due to the documented errors in the vertical temperature structure in the Arctic in ERA-40 (Bromwich and Wang 2005) we show mainly the analyses with NNR as a common reference unless specified otherwise.

c. 20 year means and trends

In order to address lower frequency variability, we analysed 20-yr averages and 20-yr trends for the seasonal and regional averages defined above. The size of the window (20-yr) reflects the fact that Arctic temperature is known to show variability on this time scale (e.g., Polyakov et al. 2003, Overland et al. 2004). The analyses are then performed with 10-yr overlapping windows (i.e., 20 yr windows moving in steps of 10 years).

The definition of start and end dates of the intervals is based on the available data. Several starting and ending years of data sets lie in the years 7-9 of a decade (NNR, ERA-40, JRA-25 and ERA-Interim start in 1948, 1957, 1979 and 1989, respectively, REC1 ends in 1957, other data sets between 2007 to 2009). Therefore, to fully exploit the lengths of the data sets we chose the intervals 1908-1927, 1918-1937, …, 1988-2007. Not more than five missing seasons are allowed; neither the first nor the last 2 years can be missing. Trends were calculated using least squares regression.

4. Results and discussion

a. Day-to-day variability in Svalbard

During the International Geophysical Year period of 1957 to 1958, three radio sonde stations were in operation in Svalbard. A mutual comparison of the simultaneous ascents from these three stations illustrates the range of differences that can be expected from nearby, simultaneous observations and from observations that are separated by distances similar to the grid spacing of the reanalysis datasets. It also provides a check on our assumed errors in radiosonde observations. The comparison is summarized in Table 2. The closest station pair (14 km distance) has smallest σ_{diff} (1.58 to 2.30 °C depending on the level). If our assumed observational error σ_{obs} of 1 °C is correct, then the error of representativeness $\sigma_{rep} = \sqrt{(\sigma_{diff}^2 - 2\ \sigma_{obs}^2)}$ = 0.7-1.8 °C. For the two station pairs that are further apart (around 240 km), σ_{diff} is larger and σ_{rep} increases to 1.9 to 2.9 °C. The differences are broadly consistent with a fixed σ_{rep} of 1.96 °C (see Table 2, lower part, and equation in section 3a), which is used elsewhere in this paper to measure errors of the interpolation of reanalyses to station locations (i.e., over distances of 0-150 km and offsets of 0-3 hours).

Mean differences reach an amplitude of 2.8 °C near the ground (note that 1000 hPa temperature is only reported if the level is above surface), largely due to real differences in temperature between the locations (after subtracting the corresponding NNR climatologies, differences decrease, cf. lower part of Table 2). Differences generally decrease at higher levels. Cape Linné (especially after subtracting the NNR climatology) shows lower temperatures in the middle troposphere than the other two stations. The difference to Barentsburg (over a distance of just 14 km) reaches 1.7 °C, pointing to a possible bias that was not detected in the quality control of

CHUAN, arguably due to the short length of the series (1.5 yrs). At 200 hPa, the mean values from all three stations (after subtracting climatology) are within 0.65 °C.

Correlations are generally above 0.75 (above 0.9 for the two closest stations) in the lower troposphere, reach a minimum near 300 hPa and then increase again to the 200 hPa level. In all, the analyses are consistent with our assumed errors. They also show, however, that there may be remaining biases in the observations that cannot be estimated easily.

In the next step we compared the station data with 20CR data interpolated to the station locations. At all three locations, 20CR shows higher temperatures than the observations at 1000 hPa (around 3 °C), slightly higher temperatures in the middle troposphere, but 10 °C lower temperatures at 200 hPa. Differences are largest compared to Cape Linné, which is likely biased cold in the observations. Correlations between 20CR and observations (after subtracting NNR climatology) reach 0.7 to 0.85 in the middle troposphere, but are lower near the ground. The fraction of the differences exceeding $\hat{\sigma}_{diff}$ is 5-15% in the middle troposphere, higher near the ground and at the tropopause level. This unexpected high exceedance rate is most likely due to the biases (if the mean difference is subtracted first, exceedance rates drop to 0.9-4.3% at all stations and all levels from 850 to 300 hPa, but remain above 5% for 1000 hPa and 200 hPa).

To expand the analysis, we compare additional available data with 20CR during a few extreme years. We analyse the cold winters 1911/1912 and 1988/1989 and the warm winters 1944/1945 and 2005/2006. The two early winters provide a particularly hard test because during these time periods, scant surface information from the Arctic was available for assimilation into 20CR.

Temperature profiles from tethered balloons and kites and from reanalysis data from Advents Bay and Ebeltofthamna, 1911-1913 are compared in Fig. 3. Values are expressed as anomalies from the daily NNR climatology (1968-1996). The observations often show strongest anomalies near the ground (note that due to a change in the reporting, no observations are available for the 200 m level after May 1912), which may be a real feature or arise from an inaccurate depiction of the surface layer in the reference (NNR). Absolute values show relatively shallow surface inversions (<200 m), and sometimes inversions at higher levels (200-1100 m asl). The profiles from 20CR (Fig. 3, linearly interpolated from pressure levels to altitude levels) are on most days much warmer near the ground (Table 3), particularly in winter and during cold days identified from the observed data. The biases are statistically significant up to 2000 m asl. The biases are very likely due to an error in specifying sea ice in 20CR,

leading to anomalous heat flux (Compo et al. 2011). However, other factors (i.e., specific local conditions, interpolation, time mismatch, etc.) might also contribute.

Despite these systematic differences, we find relatively good correlations of the anomalies on a day-to-day scale (Table 3). At the surface, correlations are low (around 0.4), but above 1000 m asl we find anomaly correlations of 0.6 to 0.8. Single warm profiles are well reproduced, but cold ones less well (both near the surface and at 1500 m asl). The differences between 20CR and observations near the ground are too frequently outside their respective errors (i.e. $P(|\Delta T|>2\hat{\sigma}_{diff}) > 0.05$) because of the warm bias near the ground. From 1000 m asl upward, however, this is the case only for 6%, which agrees well with the stated errors.

Figure 4 shows a similar analysis for the winter of 1944/1945. A warm bias at the surface in 20CR is clearly visible, and a cold tropopause bias appears (Table 4). Anomaly correlations (Table 4) are between 0.7 and 0.9 in the lower and middle troposphere. Hence, both data sets contain similar features of day-to-day variability. Strong positive temperature anomalies of 10 °C or more are represented in both data sets. However, occasionally differences between the data sets can be equally large. In terms of the fraction of differences within $\pm 2\hat{\sigma}_{diff}$, the agreement is poor at 1000 hPa (note that temperature for this level is not reported if the level is of below ground, affecting the sampling) and near the tropopause. In contrast, between the 700 hPa and the 500 hPa level, the agreement between the actual and expected differences close to that predicted (i.e., only 5.5% of the differences are outside $\pm 2\ \hat{\sigma}_{diff}$).

In view of the errors in the historical upper-air data, the interpolation procedure, and the possible effect of the time mismatch (0-3 hours) the correlations in both episodes are considered to be high. It may therefore not be surprising that we find correlations between 20CR and observations for the more recent winters 1988/1989 and 2005/2006 (not shown) are similar to the winter 1944/45, with coefficients between 0.8 and 0.9. (Note the conventional reanalyses exhibit correlations between 0.9 and 0.995 with observations for these two winters).

Biases in the recent winter 1988/89 are also similar to those find for 1944/45. 20CR is 2.3 °C warmer than observations at 850 hPa (see also Fig. 1), 0.7 °C cooler at 500 hPa and 4 °C and 12 °C cooler at 300 and 200 hPa, respectively In comparison, the conventional reanalyses are 0-3 °C cooler at 850 hPa, 0.5-1.7 °C cooler at 500 hPa, 0.9-3 °C and 0.7-2.7°C cooler at 300 and 200 hPa, respectively.

In contrast, for the winter 2005/2006, 20CR temperatures from the 850 to the 500 hPa level lie within ±0.3 °C of the observations, while larger differences are found in some cases for the conventional reanalyses. The 20CR cold bias near the tropopause remains very strong also in the winter 2005/2006. The improvement in the low-level comparison may be a result of actual reduced sea ice concentrations near Svalbard (Cottier et al. 2007) ameliorating the impact of the 20CR coastal misspecification of sea ice concentration. This would suggests that future historical reanalyses may have a substantial reduction in their lower tropospheric Arctic biases compared to 20CR.

In summary, the analyses of cold and warm winters shows that day-to-day temperature variability is rather well captured in 20CR between about 850 hPa and 500 hPa. There are systematic differences near the ground and near the tropopause. In the 1911/12 case the agreement is better for warm days than cold days. Overall, where 20CR biases are small (i.e., the middle troposphere), actual and expected differences are consistent and the variability in observations and reanalyses is similar.

b. Interannual variability

After addressing four specific winters in Svalbard, we next compare seasonal mean values from Svalbard for the period 1908 to 2007 (Fig. 1). The warmest winter was 2005/2006 (both at the surface and at 850 hPa; only in NNR 2006/07 was slightly warmer at 850 hPa). The coldest winters were 1916/17 (surface), 1917/18 (REC1 at 850 hPa), and 1962/63 (all other data sets). For 850 hPa temperature, correlations between observations and gridded products, over the entire respective periods, are on the order of 0.9 for conventional reanalyses (which include these observations) and 0.8 for 20CR (which is independent). During the most recent period 1989-2008, correlations with observations are ~0.95 for all gridded data sets (20CR, NNR, ERA-Interim, JRA-25). Hence, Svalbard's interannual variability is relatively well captured (REC1 and REC2 have too little overlap with observations).

For a more comprehensive examination of interannual variability around the Arctic, Figs. 5-8 show seasonal-regional averages for all data sets in the form of time-height cross-sections. The plots provide a useful visual tool for detecting different characteristics of data sets. They allow one to address even subtle details. Quantitative results are given in tables and supplementary

material. We use NNR as reference climatology here for all data sets. Note that, in several cases, we have combined more than one data set in one panel for ease of presentation.

Examining Figs. 5-8, there are obvious differences between the data sets in terms of absolute values of the anomalies. Starting at near the surface, 20CR is warmer than NNR and in fact warmer than all other data sets.. ERA-40 and ERA-Interim are also warmer near the surface than NNR. This is probably due to an error in NNR. Both ERA datasets and 20CR have prescribed fractional sea ice concentration in a grid box, while NNR has prescribed either 100% or 0% only. Such a specification results in too little heat flux from the ocean to the atmosphere when fractional sea ice is present. Note that in the case of 20CR part of the difference near the surface can be attributed to an error in the specification of the sea-ice concentration (Compo et al. 2011). However, other factors including the representation of orography and the interpolation to pressure levels might also contribute.

While there are interesting variations throughout the troposphere, the most noticeable issue is a cold bias in 20CR near the tropopause compared to the other datasets. This bias is not constant over time but increases strongly in the 1930s and 1940s. The cause of this bias and its variability is unknown.

Looking at these two issues in 20CR in more detail, compared with observations, the surface warm bias is largest in winter and spring (see Table 5 and Supplementary tables), with large regional differences. The largest biases are found over the Canadian Arctic and the smallest biases are found over the European Arctic.The cold tropopause bias has a similar seasonal and regional distribution.. There is a negative correlation of the two errors on an interannual scale, i.e., years with a strong surface warm bias also tend to have a strong tropopause cold bias, which for some seasons and regions is statistically significant.

Returning to the broader comparison of the several data sets, the amount of variability varies greatly between them. CHUAN shows a relatively high variability in the early years that contrasts with that in later years. This increased variability is very likely an artifact of the sparse sampling of the upper-air stations. Conversely, REC1 or REC2 show very little variability, which is understandable as they are based on linear regression and thus underestimate the variance by construction. The 20CR shows a similar amount of variability in the earlier period as in later periods and, for the free troposphere, is similar to the other reanalysis data sets.

Several multiannual features in mid-tropospheric temperature appear in all data sets, e.g., the cold winters in 1940-1942 in northeastern Europe that extended into the Arctic sector. These wintertime anomalies were likely related to an El Niño event (Brönnimann et al. 2004; also see Brönnimann 2007 for a general discussion of El Niño effects on Europe). Also noticeable are the warm winters in the early 1970s (Fig. 5). Prominent multiannual features in other regions and seasons are the cold anomalies in spring in Western Siberia in the 1960s (Fig. 6) and the warm 1990s in almost all seasons and sectors (Figs. 5-8). The warm anomalies in the NNR in the upper troposphere over Western Siberia in spring in the late 1940s and early 1950s (Fig. 6) are to some degree attributable to errors in data processing. Approximately 30 stations in the former Soviet Union have a suspected undercorrected radiation and lag error during that period, which is corrected in CHUAN but not in NNR (Grant et al. 2009a).

The interannual variability is similar in most data sets. As an example, Table 5 shows the correlation between 20CR and CHUAN in the European Arctic in all seasons. Correlations are between 0.75 and 0.93 for winter and autumn throughout the lower and middle troposphere. Correlations decrease at the tropopause level (due to varying tropopause height), and they are smaller for the spring and summer seasons. Examining correlations to the 1930-1957 period allows all historical data sets to be compared (Table 6 for Dec.-Feb.). REC1 shows the lowest correlation with observations (CHUAN) as well as with other data sets (REC2, 20CR). In the lower troposphere, the highest correlations are found for 20CR. REC2 shows slightly lower correlations with CHUAN than 20CR in the lower troposphere but shows the highest correlations of all data sets in the upper troposphere and stratosphere.

Expanding the comparison to the more recent period and including the conventional reanalyses leads to a similar conclusion that the interannual variability is very similar in the several datasets. In Table 7, monthly anomalies from each dataset's own climatology for the European Arctic region are compared. The climatology is changed to avoid seasonally dependent biases in the NNR climatology. 20CR and REC2 both show high correlations with each other and with NNR in the lower and middle troposphere. ERA40 and 20CR also compare well in the free troposphere.

Corresponding tables for the other sectors are given in the supplementary material. Because of the lower amount of available observations (CHUAN), the correlations vary more strongly, but support the results seen inTables 5 and 7.

The main result of this comparison is that all data sets agree well among each other with respect to interannual variability. REC1 agrees slightly less well with the other data sets, 20CR agrees well in the troposphere but not the stratosphere, while REC2 agrees well also in the upper troposphere and stratosphere.

c. Bi-decadal means and trends

We now analyse trends and mean values over longer time periods. We first return to the long record of 850 hPa temperature in winter over Svalbard (Fig. 1). Although interannual variability was relatively similar comparing the datasets, there are substantial differences even in the relatively recent 1980-2002 trend from ERA40 (0.85 °C/decade), NNR (0.57 °C/decade), JRA25 (0.49 °C/decade), observations from Ny Ålesund and Barentsburg merged (0.35 °C/decade), and 20CR (0.19 °C/decade). These large discrepancies among the data sets underscore the large uncertainties involved with estimates of the trend.

For the seasonal-regional averages, Figs. 9-12 show vertical structures of temperature trends in overlapping 20-yr periods for different data sets. Trends are not consistent through time, space, and season. Positive trends alternate with negative trends, though it is visually apparent that positive trends dominate in the troposphere compared to negative trends in the lower stratosphere.

A common feature seen in Figs. 9 and 10 is that the tropospheric warming is especially strong in the 1978-1997 period. The Canadian Arctic shows the strongest warming for themost recent period. A further common feature is the cooling trend throughout the troposphere in 1948-1967 in almost seasons and regions (except in 20CR for the Canadian Arctic (Fig. 12)). In NNR over Siberia in spring (Fig. 10), the more pronounced cooling is very likely due to the warm bias in the first half of that period. However, other data sets also show a consistent cooling.

Concerning the vertical structure, almost all recent warming trends (1978-1997 and 1988-2007), with the most notable exception of the summer trend in the Pacific Arctic, are stronger near the ground than at 700 hPa. The structure of the trend during the ETCW (1918-1937) is less clear. In 20CR it is also stronger near the ground than at 700 hPa.

We find the following trend differences between the data sets:

- For the European Arctic in winter (Fig. 9), CHUAN shows a more pronounced warming at 700 hPa from 1938 to 1957 than 20CR, REC1, or REC2, while the cooling from 1928 to 1947 at this level is more pronounced in REC1 than in REC2 or 20CR (CHUAN has insufficient data).

- In the Siberian Arctic in spring (Fig. 10), the lower tropospheric warming from 1958 to 1977 is weaker in 20CR and ERA-40 than in CHUAN or NNR.

- In the Pacific Arctic in summer (Fig. 11), the sign of the trend in the lower troposphere does not agree between 20CR and REC1 from 1928 to 1947 or between 20CR and CHUAN in 1948 to 1967.

- In the eastern Canadian region in fall (Fig. 12), 20CR (JRA-25) shows a weaker tropospheric warming over the period 1988-2007 (1978-1997) than all other data sets. 20CR and REC1 disagree in the sign of the tropospheric trend throughout the first half of the twentieth century.

While the differences between CHUAN, 20CR, and reconstructions are understandable from the relatively large differences in their input data and their approaches, the differences between the more conventional reanalyses must be related to other factors such as the changes in the assimilation systems, data processing, or in the observation network.

The 20-yr trends show large differences from one time window to the next. In order to focus on the multidecadal changes, we compare 20-yr averages for these different time windows in Figs. 13 and 14. Here the data are expressed with respect to self-climatologies of the period 1961-1990 in order to remove biases (consequently, JRA-25 and ERA-Interim cannot be shown). First we focus on a comparison between the warm periods 1918-1937 (only 20CR, REC1, and CRUTEM3v are available for this period) and 1988-2007 (CHUAN, 20CR, NNR), respectively (Fig. 13). The profiles are well constrained in the recent period, while there are relatively large differences in 1918-1937. However, despite these differences a change in the profile shape appears in the sense that lower tropospheric lapse rates are larger in 1988-2007 in most data sets and seasons compared to 1918-1937.

In order to extend the analysis to all 20-yr periods in all data sets, in Fig. 14 we concentrate on the average and range of all available observation-based data sets (including CHUAN) to highlight common features. Care should be taken in the interpretation of such an “ensemble”

mean. Most or all observation-based data have issues in the Arctic that may affect the trends in the vertical structure. (A comprehensive version of the figure with each data set shown as a different symbol is given in the supplementary material).

The range in the ensemble of observation-based data sets for the early twentieth century is affected by likely artificial trends in 20CR. Most notably, 20CR shows much higher anomalies than the other data sets at 200 hPa in autumn to spring and the opposite near the ground in summer. Figure 14 shows that the average for the last 20-yr period (1988-2007) not only differs from the 1918-1937 period, but from all other periods. The average over all data sets (solid line) is outside the range (bars) of any period in all seasons up to 850 hPa, in some seasons higher. For the summer and fall study regions, the range for 1988-2007 does not overlap with the range for any previous period in the lower troposphere.

The only instance where the 1918-1937 warm period rivals the recent anomaly concerns temperature at 700 hPa and higher levels in winter in the European Arctic. Grant et al. (2009a) found a very strong coincidence of this warm anomaly with anomalous meridional advection from central Europe to the European Arctic in REC1. This is also confirmed by all other data sets discussed here (not shown). They also found this advection to be consistent with western European sulphate aerosols deposited in a Svalbard ice core.

Note that we are comparing the 1918-1937 period with the 1961-1990 average. The conclusion might be different when comparing to the 1910s. Isaksson et al. (2005), based on ice core data from different elevations and comparison with early station data, suggest that the cold period prior to the 1920s at Svalbard was due to more frequent inversions. Indeed, Fig. 1 suggests that an abrupt shift around 1918/1919 was much larger near the ground than at 850 hPa.

5. Conclusions

Different observation-based data sets were analysed with respect to their ability to represent the vertical thermal structure of the Arctic troposphere on different time scales. The analyses revealed excellent agreement in terms of correlation at various time scales, but they also revealed several inaccuracies in the four long data sets that cover the ETCW. 20CR has a warm bias near the ground due to misspecification of sea-ice (Compo et al. 2011) that regionally and seasonally can exceed 10 °C. Moreover, there is a cold bias near the tropopause, which

increases in the 1930s and 1940s and also exceeds 10 °C. Upper-air observations may have remaining instrumental biases that are difficult to quantify, especially in the early years. Furthermore, regional averages constructed from the data exhibit too much variability (which could be remedied using a variance correction). Finally, by construction REC1 and REC2 have too little variability and have little skill in summer at stratospheric levels. The validation statistics of REC2 (Brönnimann et al. 2011a) indicate a higher skill than REC1, but point to systematic deficiencies in the Russian Arctic. Both reconstructions have not been validated for trend analysis.

These problems add to the list of known shortcomings of the conventional reanalysis data sets. ERA-40 has problems with satellite radiance assimilation over the ice-covered Arctic Ocean (Bromwich and Wang 2005, Uppala et al. 2005), as discussed above. Bromwich et al. (2007) performed an assessment for the conventional reanalyses ERA-40, NNR, and JRA-25 in the polar regions and discussed differences in the data sets (addressing also clouds and cyclones). Lüpkes et al. (2010) compared ERA-Interim data with ship-based observations and found problems related to sea-ice in ERA-Interim. NNR has a warm bias over the former Soviet Union in 1948–1957 due to uncorrected radiation errors in the radiosonde data (Grant et al. 2009a).

Based on our comparisons we conclude that synoptic scale variability is best analysed in 20CR (or CHUAN, if data is available), provided that the biases are taken into account. Interannual variability is similarly well represented in all four data sets (20CR, REC1, REC2, and CHUAN), apart from difference in the mean and in the variance. Hence, for correlation analyses with other variables, all data sets can be used. Among the datasets, REC2 has the highest correlations with observations at the 300 hPa and 200 hPa levels, but is not spatially complete.

None of the data sets alone is sufficient for addressing long-term trends in the Arctic. However, knowing the shortcomings and differences, information can be gained even on trends from analysing all data sets individually and by combining the results (see also Thorne et al. 2010 for the value of multiple tropospheric temperature data sets). For instance, all data sets agree that the last two decades are unprecedented in the 20th century in terms of the magnitude of the warm anomaly in the lower troposphere. The rate of warming between the 1980s and present is also outstanding. The vertical structure of the trend shows a clear amplification of the recent trend at the surface in autumn to spring. During the ETCW, high temperature anomalies were

also found at 700 hPa and above in winter. Although the data are more uncertain for the first half of the twentieth century, they clearly point to a smaller lapse rate compared to the recent warm period.

Acknowledgment. This work was supported by the Swiss National Science Foundation through the project "Past climate variability from an upper-level perspective" as well as through the National Competence Centre in Research (NCCR) Climate. The Twentieth Century Reanalysis Project (20CR) used resources of the National Energy Research Scientific Computing Center and of the National Center for Computational Sciences at Oak Ridge National Laboratory, which are supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and Contract No. DE-AC05-00OR22725, respectively. Support for the 20CR dataset is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration Climate Program Office. Constructive comments by two anonymous reviewers and also by P.D. Sardeshmukh of the University of Colorado CIRES/CDC and NOAA/ESRL/PSD on an earlier version of this manuscript are greatly appreciated.

References

Bengtsson L, Semenov V, Johannessen O (2004) The early twentieth-century warming in the Arctic–a possible mechanism. J Clim 17:4045–4057

Bitz CM, Fu Q (2008) Arctic warming aloft is data set dependent. Nature 455:E3-E4

Brohan P, Kennedy J, Harris I, Tett S, Jones P (2006) Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. J Geophys Res 111: D12106. doi:10.1029/2005JD006548.

Bromwich D, Wang S (2005) Evaluation of the NCEP-NCAR and ECMWF 15-and 40-yr reanalyses using rawindsonde data from two independent Arctic field experiments. Mon Wea Rev 133:3562–3578

Bromwich DH, Fogt RL, Hodges KI, Walsh JE (2007) A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. J Geophys Res 112:D10111, doi:10.1029/2006JD007859

Brönnimann S (2003) A historical upper-air data set for the 1939-1944 period. Int J Climatol 23:769-791

Brönnimann S (2009) Early twentieth-century warming. Nature Geosc 2:735-736

Brönnimann S, Luterbacher J, Staehelin J, Svendby TM, Hansen G, Svenøe T (2004) Extreme climate of the global troposphere and stratosphere in 1940-42 related to El Nino. Nature 431:971-974

Brönnimann S, Griesser T, Stickler, A (2011a) A gridded monthly upper-air data set back to 1918. Clim Dyn doi:10.1007/s00382-010-0940-x (online first)

Brönnimann S, Compo G.P, Allan R, Adam W, Spadin R (2011b) Early ship-based upper-air data and comparison with the Twentieth Century Reanalysis. Clim Past 7:265-276, doi:10.5194/cp-7-265-2011

Cottier FR, Nilsen F, Inall ME, Gerland S, Tverberg V, Svendsen H (2007) Wintertime warming of an Arctic shelf in response to large-scale atmospheric circulation. Geophys Res Lett 34:L10607, doi:10.1029/2007GL029948

Compo GP and co-authors (2011) The Twentieth Century Reanalysis Project. Q J R Meteorol Soc 137:1-28, doi: 10.1002/qj.776

Dee DP and co-authors (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137, 553–597. doi: 10.1002/qj.828

Dege W (1960) Wissenschafliche Beobachtungen auf dem Nordostland von Spitzbergen. Berichte des Deutschen Wetterdienstes Nr. 72 (Bd. 10), Offenbach a M

Durre I, Vose RS, Wuertz DB (2006) Overview of the Integrated Global Radiosonde Archive. J Clim 19:53-68

Grant A, Brönnimann S, Haimberger L (2008) Recent Arctic warming vertical structure contested. Nature 455:E2-E3, doi:10.1038/nature07257

Grant A, Brönnimann S, Ewen T, Nagurny A (2009a) A new look at radiosonde data prior to 1958. J Clim 22:3232-3247

Grant AN, Brönnimann S, Ewen T, Griesser T, Stickler A (2009b) The early twentieth century warm period in the European Arctic. Met Z 18:425-432

Graversen R, Mauritsen T, Tjernström M, Källèn E, Svensson G (2008) Vertical struture of recent Arctic warming. Nature 451:53–56

Griesser T, Brönnimann S, Grant A, Ewen T, Stickler A, Comeaux J (2010) Reconstruction of global monthly upper-level temperature and geopotential height fields back to 1880. J Clim 23:5590-5609

Haimberger L (2007) Homogenization of radiosonde temperature time series using innovation statistics. J Clim 20: 1377–1403

Hansen J, Ruedy R, Glascoe J, Sato M (1999) GISS analysis of surface temperature change. J Geophys Res 104:30997-31022

Isaksson E, Kohler J, Pohjola V, Moore J, Igarashi M, Karlof L, Martma T, Meijer H, Motoyama H, Vaikmae R, van de Wal RSW (2005) Two ice-core delta O-18 records from Svalbard illustrating climate and sea-ice variability over the last 400 years. The Holocene 15:501-509

Kauker F, Köberle C, Gerdes R, Karcher M (2008) Modeling the 20th century Arctic Ocean/Sea Ice system: Reconstruction of surface forcing. J Geophys Res 113:C09027

Kistler R and co-authors (2001) The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. B Amer Meteorol Soc 82:247-267

Johannessen OR and co-authors (2004) Arctic climate change: observed and modelled temperature and sea-ice variability. Tellus 56A:328–341

Lüpkes C, Vihma T, Jakobson E, König-Langlo G, Tetzlaff A (2010) Meteorological observations from ship cruises during summer to the central Arctic: A comparison with reanalysis data. Geophys Res Lett 37:L09810, doi:10.1029/2010GL042724

Onogi K and co-authors (2007) The JRA-25 Reanalysis. J Met Soc Jap 85:369-432

Overland J, Spillane M, Percival D, Wang M, Mofjeld H (2004) Seasonal and regional variation of Pan-Arctic surface air temperature over the instrumental record. J Clim 17:3263–3282

Petoukhov V, Semenov VA (2010), A link between reduced Barents-Kara sea ice and cold winter extremes over northern continents. J Geophys Res 115:D21111

Polyakov I, Bekryaev R, Alekseev G, Bhatt U, Colony R, Johnson M, Maskshtas A, Walsh D (2003) Variability and trends of air temperature and pressure in the maritime Arctic, 1875-2000. J Clim 16:2067–2077

Przybylak R (2007) Recent air-temperature changes in the Arctic. Annals of Glaciology 46:316-324

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late Nineteenth Century. J Geophys Res 108:4407, doi:10.1029/2002JD002670

Rempp G, Wagner A (1916) Die Temperaturverhältnisse über Spitzbergen in der Adventbai 1911/12. Veröffentlichungen des Deutschen Observatoriums Ebeltofthafen-Spitzbergen 1, 27 pp.

Screen JA, Simmonds I (2010) The central role of diminishing sea ice in recent Arctic temperature amplification. Nature 464:1334–1337

Screen JA, Simmonds I (2011) Erroneous Arctic temperature trends in the ERA-40 reanalysis: a closer look. J Clim 24:2620-2627

Selinger F (2001) Von „Nanok“ bis „Eismitte“ - Meteorologische Unternehmungen in der Arktis 1940-1945. Schriften des Deutschen Schiffahrtsmuseums

Serreze MC, Barrett AP, Stroeve JC, Kindig DN, Holland MM (2009) The emergence of surface-based Arctic amplification. The Cryosphere 3:11-19

Stickler A and co-authors (2010) The Comprehensive Historical Upper Air Network (CHUAN). B Amer Meteorol Soc 91:741-751. doi: 10.1175/2009BAMS2852.1

Thorne P (2008)Arctic tropospheric warming amplification? Nature 455:E1-E2, doi:10.1038/nature07256

Thorne, PW, JR Lanzante, TC Peterson, DJ Seidel, KP Shine (2010) Tropospheric temperature trends: history of an ongoing controversy. Advanced Review 2:66-8

Tuomenvirta H, Drebs A, Førland E, Tveito OE, Alexandersson H, Vaarby Laursen E, Jónsson T (2001) Nordklim data set 1.0 - description and illustrations. Norwegian Meteorological Institute Report 08/01 Klima, Oslo

Uppala SM and co-authors (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131:2961-3012

Wegener K (1916) Die Technik der Drachen- und Ballonaufstiege im Winter 1912/13 zu Ebeltofthafen (Spitzbergen). Veröffentlichungen des Deutschen Observatoriums Ebeltofthafen-Spitzbergen, Heft 2, 3-9

Wegener K, Robitzsch M (1916a) Ergebnisse der Pilot-Visierungen während der Überwinterung 1912/13. Veröffentlichungen des Deutschen Observatoriums Ebeltofthafen-Spitzbergen 3, 18 pp

Wegener K, Robitzsch M (1916b) Ergebnisse der Fessel-Aufstiege während der Überwinterung 1912/13. Veröffentlichungen des Deutschen Observatoriums Ebeltofthafen-Spitzbergen 4, 21 pp

Wang M, Overland J, Kattsov V, Walsh J, Zhang X, Pavlova T (2007) Intrinsic versus forced variation in coupled climate model simulations over the Arctic during the twentieth century. J Clim 20:1093–1107

Wood KR, Overland JE (2010) Early 20th century Arctic warming in retrospect. Int J Climatol 30:1269-1279, doi: 10.1002/joc.1973

Wood KR, Overland JE, Jónsson T, Smoliak BV (2010) Air temperature variations on the Atlantic‐Arctic boundary since 1802. Geophys Res Lett 37:L17708, doi:10.1029/2010GL044176

Table 1. Upper-air data sets used in this study. Note that time period, time resolution, and spatial resolution represent the form in which the data sets were used in this study, not the original resolutions and time periods. UA = upper-air observations, SLP = sea-level pressure, SAT = surface air temperature, SST = sea-surface temperature

#	Data set	Abbr.	Period	Type	Input	Time resolution	Spatial resolution	Reference
1	Comprehensive historical upper-air network	CHUAN	1930[+]-2006	Observations	-	State	135 Arctic stations	Stickler et al. 2010 Grant et al. 2009 Brönnimann 2003
2	Reconstructions	REC1	1880-1957	Statistical reconstructions	UA, SLP, SAT*	monthly mean	2.5°	Griesser et al. 2010
3	Reconstructions	REC2	1923-2001	Statistical reconstructions	UA, SLP, SAT*	monthly mean	2.5°	Brönnimann et al. 2011a
4	Twentieth century reanalysis, vers. 2	20CR	1871-2008	Data assimilation (Ensemble Kalman Filter, NCEP/GFS model)	SLP, monthly SST	6-hourly	2°	Compo et al. 2011
5	NCEP/NCAR reanalysis	NNR	1948-2009	Data assimilation (Statistical Interpolation, NCEP/MRF model)	All	Daily	2.5°	Kistler et al. 2001
6	European reanalysis	ERA-40	1957-2002	Data assimilation (3D Var, IFS model)	All	Monthly	2.5°	Uppala et al. 2005
7	Japanese reanalysis	JRA-25	1978-2008	Data assimilation (3D-Var, JMA model)	All	Monthly	2.5°	Onogi et al. 2007
8	European reanalysis	ERA-Interim	1989-2009	Data assimilation (4D Var, IFS model)	All	Monthly	1.5°	-

* ERA-40 was used for calibration

[+] except the record from Advents Bay/Ebeltofthamna (1911-1913)

Table 2. Comparison of temperatures from three stations at Svalbard during the International Geophysical Year 1957/58: *n* is the number of paired observations, ΔT is the averaged difference between two records. Based on *n* the standard deviation (σ_{diff}) and the correlation coefficient (r) is given. $P(|\Delta T|>2\ \hat{\sigma}_{diff})$ is the fraction of differences outside the interval $\pm 2\ \hat{\sigma}_{diff}$ estimated form assuming σ_{obs} = 1 °C and σ_{rep} = 1.96 °C. The upper part of the table shows the comparison of the raw data, the middle part shows the results for individual ascents minus a daily NNR climatology, 1968-1996, linearly interpolated to the observations. The lower part shows comparisons between the station observations and 20CR at the grid point 16° E, 80° N.

	Cape Linné minus Kinnvika 242 km distance			Cape Linné minus Barentsburg 14 km distance			Kinnvika minus Barents 238 km distance		
pressure (hPa)	*n*	ΔT (°C)	σ_{diff} (°C)	*n*	ΔT (°C)	σ_{diff} (°C)	*n*	ΔT (°C)	σ_{diff}
1000	182	2.30	3.00	181	-0.46	1.87	180	-2.77	3.
850	225	0.90	3.22	227	-0.67	1.65	225	-1.57	3.
700	223	0.00	2.60	226	-1.25	1.58	224	-1.26	2.
500	221	-0.81	2.74	222	-1.61	1.67	222	-0.80	3.
400	219	-0.96	2.85	218	-1.73	2.25	218	-0.73	3.
300	218	-0.82	2.54	207	-1.41	2.30	213	-0.53	2.
200	208	-0.84	2.40	175	-0.64	2.19	184	0.23	2.
pressure (hPa)	*r*	ΔT (°C)	$P(\lvert\Delta T\rvert>2\ \hat{\sigma}_{diff})$	*r*	ΔT (°C)	$P(\lvert\Delta T\rvert>2\ \hat{\sigma}_{diff})$	*r*	ΔT (°C)	$P(\lvert\Delta T\rvert>$
1000	0.833	-1.65	0.165	0.912	-0.47	0.017	0.732	1.14	0.1
850	0.824	-0.50	0.076	0.940	-0.67	0.009	0.847	-0.17	0.0
700	0.851	-0.99	0.045	0.941	-1.25	0.009	0.813	-0.28	0.0
500	0.824	-1.56	0.054	0.932	-1.61	0.023	0.789	-0.06	0.0
400	0.763	-1.59	0.082	0.861	-1.73	0.041	0.743	-0.11	0.0
300	0.697	-1.08	0.046	0.772	-1.42	0.048	0.617	-0.28	0.0
200	0.876	-0.65	0.034	0.892	-0.65	0.029	0.855	0.00	0.0

pressure (hPa)	20CR minus Cape Linné			20CR minus Barentsburg			20CR minus Kinnvik		
	r	ΔT (°C)	$P(\|\Delta T\|>2\ \hat{\sigma}_{diff})$	*r*	ΔT (°C)	$P(\|\Delta T\|>2\ \hat{\sigma}_{diff})$	*r*	ΔT (°C)	$P(\|\Delta T\|>$
1000	0.647	3.42	0.250	0.482	2.94	0.279	0.752	3.32	0.2
850	0.810	1.28	0.101	0.835	0.61	0.088	0.811	1.18	0.1
700	0.847	1.56	0.058	0.824	0.30	0.075	0.845	0.77	0.0
500	0.793	2.00	0.112	0.756	0.41	0.120	0.762	0.63	0.0
400	0.703	1.45	0.121	0.695	-0.27	0.144	0.651	0.04	0.1
300	0.335	-0.58	0.177	0.310	-1.91	0.327	0.362	-1.69	0.2
200	0.282	-9.97	0.755	0.320	-10.78	0.834	0.272	-10.84	0.7

Table 3. Comparison of temperature anomalies between upper-air observations (individual ascents minus a daily NNR climatology, 1968-1996, linearly interpolated to the observations) and 20CR (closest standard time, linearly interpolated to the observations) for Svalbard, 1911-1913. *n* is the number of paired observations, ΔT is the averaged difference between 20CR and observations, *r* is the correlation coefficient, and $P(|\Delta T|>2\ \hat{\sigma}_{diff})$ is the fraction of differences outside the interval $\pm 2\sigma_{diff}$. Bold numbers indicate differences that are significantly different from zero (two sided t-test, $p<0.05$).

altitude (m asl)	200	500	1000	1500	2000	2500	3000	3500
n	78	165	125	80	39	19	11	7
r	0.405	0.397	0.514	0.600	0.585	0.686	0.846	0.925
ΔT (°C)	**8.4**	**3.2**	**2.4**	**0.5**	**0.8**	0.8	0.6	-0.7
$P(\|\Delta T\|>2\ \hat{\sigma}_{diff})$	0.590	0.194	0.027	0.088	0.026	0	0	0

Table 4. Comparison of temperature anomalies between upper-air observations (individual ascents minus a daily NNR climatology, 1968-1996, linearly interpolated to the observations) and 20CR (closest standard time, linearly interpolated to the observations) for Svalbard, 1944-1945. *n* is the number of paired observations, ΔT is the averaged difference between 20CR and observations, *r* is the correlation coefficient, and $P(|\Delta T|>2\ \hat{\sigma}_{diff})$ is the fraction of differences outside the interval $\pm 2\sigma_{diff}$. Note that the 1000 hPa level is affected by a sampling bias in that observations are only available if the level was above the Earth's surface. All differences are significantly different from zero (two sided t-test, $p<0.05$).

pressure level	1000 hPa	850 hPa	700 hPa	600 hPa	500 hPa	400 hPa	300 hPa	200 hPa
n	95	132	132	132	132	131	120	106
r	0.729	0.813	0.875	0.877	0.863	0.800	0.515	0.442
ΔT (°C)	4.7	1.2	0.7	0.7	0.6	-0.2	-2.6	-11.2
$P(\|\Delta T\|>2\ \hat{\sigma}_{diff})$	0.821	0.092	0.053	0.045	0.068	0.126	0.181	0.274

Table 5: Comparison between seasonal mean temperatures of 20CR and CHUAN for the European Arctic for different levels (note that SAT from CRUTem3v is used instead of CHUAN 1000 hPa temperature). *n* gives the number of seasonal means used for the analysis, *r* is the correlation coefficient, and ΔT is the averaged difference between 20CR and CHUAN. Correlations >0.75 are in bold. All differences are significantly different from zero (two sided t-test, $p<0.05$) except for DJF, 700 hPa and 400 hPa and SON, 850 and 700 hPa, respectively. Note the drop in *n* at 400 hPa due to the reporting in CHUAN.

		1000 hPa	850 hPa	700 hPa	500 hPa	400 hPa	300 hPa	200 hPa
n	DJF	97	54	66	62	51	59	59
	MAM	97	54	64	61	50	61	61
	JJA	97	53	64	61	51	60	60

	SON	96	54	65	62	51	59	59
r	DJF	**0.852**	**0.885**	**0.828**	**0.754**	0.714	0.472	0.472
	MAM	**0.873**	**0.870**	**0.757**	0.706	0.557	0.074	0.074
	JJA	**0.907**	**0.805**	0.657	0.368	0.518	0.271	0.271
	SON	**0.931**	**0.918**	**0.818**	**0.883**	**0.900**	0.292	0.292
ΔT (°C)	DJF	2.46	0.26	-0.12	0.24	-0.09	-1.20	-7.24
	MAM	2.38	1.03	0.75	1.06	0.82	-1.29	-10.69
	JJA	1.51	1.25	1.04	1.68	1.84	1.44	-6.16
	SON	1.25	0.05	-0.08	0.74	0.84	0.67	-4.90

Table 6: Correlations between Dec.-Feb. mean temperatures for the European Arctic for different levels in 20CR, CHUAN, REC1 and REC2 for the period 1930-1957. Correlations >0.75 are in bold (*n* = 28 except for CHUAN).

Comparison	850 hPa	700 hPa	500 hPa	300 hPa	200 hPa
n (CHUAN)	9	21	17	15	14
CHUAN-20CR°	**0.981**	**0.795**	0.700	0.124	0.174
CHUAN-REC2^{+}	**0.912**	0.714	0.702	0.742	**0.806**
CHUAN-REC1^{+}	0.714	0.619	0.543	0.729	0.650
REC1-REC2^{+}*§	**0.762**	**0.772**	0.721	0.619	0.675
REC1-20CR*	**0.825**	**0.808**	**0.752**	0.553	0.013
20CR-REC2*	**0.905**	**0.915**	**0.845**	0.368	0.070

° fully independent data sets

$^{+}$ data sets share some of the upper-air input data

* data sets share some of the SLP input data

§ data sets share the methodological approach

Table 7: Correlations of monthly temperature anomalies (with respect to the period 1961-1990 in each data set; ERA-40 was used for REC1 and REC2) for the European Arctic between different gridded data sets. REC2 has 16 missing values; all other records are complete.

Comparison	Period	850 hPa	700 hPa	500 hPa	300 hPa	200 hPa
20CR-REC1*	1923-1957	0.792	0.788	0.742	0.362	0.196
20CR-REC2*	1923-1957	0.873	0.873	0.836	0.421	0.246
REC1-REC2^{+}*§	1923-1957	0.850	0.854	0.833	0.641	0.586
20CR-NNR*§	1948-1957	0.934	0.939	0.928	0.686	0.464
REC1-NNR^{+}*	1948-1957	0.834	0.848	0.828	0.721	0.720
REC2-NNR^{+}*	1948-1957	0.939	0.951	0.951	0.897	0.791
20CR-NNR*§	1958-2001	0.941	0.961	0.947	0.637	0.369
20CR-ERA40$^{+\S}$	1958-2001	0.947	0.959	0.932	0.658	0.369
ERA40-NNR*§	1958-2001	0.986	0.987	0.978	0.904	0.985

$^{+}$ data sets share some of the upper-air input data

* data sets share some of the SLP input data

§ data sets share the methodological approach

Figure captions

Fig. 1: Svalbard time series of winter (Dec.-Feb.) of temperature averages at 850 hPa from all available time series (top) as well as surface air temperature from Svalbard assembled by the NORDKLIM project (bottom). Coloured bars indicate the time period covered by the individual data sets, grey bars indicate the winters studied in Section 4a. All series were adjusted to the location of Barentsburg for comparison, using a 1968-1996 climatology from NNR. The locations of the stations Ebeltofthamna (E), Nordaustlandet (H), Barentsburg (B), Ny Ålesund (N), Kinnvika (K) and Cape Linné (C) are indicated in Fig. 2. Another long series (not included here) is available from Bjørnsøja (Ø in Fig. 2), further to the south. For the calculation of seasonal mean values from observations see Sect. 3c.

Fig. 2: Map showing the upper-air stations in the Arctic used in this study along with the equal area grid cells used for regional averaging and the four regions for which analyses are presented. The colour indicates the start year of the record.

Fig. 3: Anomalies of daily temperature profiles (as a function of altitude above msl) from Svalbard, Nov. 1911-May 1912, Jul. 1912-Sep. 1912, Apr. 1913-Jul. 1913, from observations (top), 20CR (middle, both with respect to a 1968-1996 climatology from NNR), and their difference (bottom).

Fig. 4: Anomalies of daily temperature profiles (as a function of pressure) from Svalbard, 1944-1945, from observations (left), 20CR (middle), and their difference (right). Anomalies are constructed as in Fig. 3. Because of differences in reporting (925 hPa in NNR, 900 hPa in observations and 20CR), no climatology and hence no anomalies are available for 925 hPa.

Fig. 5: Time-height cross-section of seasonal mean temperature anomalies as a function of pressure and time for different data sets for the European Arctic (see Fig. 2) in winter. All anomalies are with respect to NNR (1961-1990) except CRUTEM3v (self-climatology, see Brohan et al. 2006). Note that for visualisation purposes, non-overlapping data sets have been combined in some cases, indicated by dashed lines). Between the end of the reconstruction period of REC2 (1957) and the start of ERA-Interim (1989) we show the calibration period of REC2. Yellow colours denote missing values.

Fig. 6: Same as Fig. 5 for Western Siberia (see Fig. 2) in spring.

Fig. 7: Same as Fig. 5 for the Pacific Arctic region (see Fig. 2) in summer.

Fig. 8: Same as Fig. 5 for the eastern Canadian Arctic (see Fig. 2) in autumn.

Fig. 9: Trend in seasonally-averaged temperature profiles over 20-yr periods as a function of pressure and time period for different data sets for the European Arctic (see Fig. 2) in winter. Note that for visualisation purposes, non-overlapping data sets have been combined in some cases, indicated by dashed lines). Between the end of the reconstruction period of REC2 (1957) and the start of ERA-Interim (1989) we show the calibration period of REC2. Yellow colours denote missing values.

Fig. 10: Same as Fig. 9 for Western Siberia (see Fig. 2) in spring.

Fig. 11: Same as Fig. 9 for the Pacific Arctic region (see Fig. 2) in summer.

Fig. 12: Same as Fig. 9 for the eastern Canadian Arctic (see Fig. 2) in autumn.

Fig. 13: Temperature anomaly averages (relative to self-climatologies 1961-90) in two 20-yr windows for different data sets for different seasonal-regional averages (a = European Arctic in winter, b = Western Siberian Arctic in spring, c = Pacific Arctic in summer, d = eastern Canadian Arctic in autumn). Blue symbols and dashed lines denote 1918-1937, red lines and symbols denote 1988-2007. Note that the latter two sectors have insufficient surface temperature data in 1918-1937.

Fig. 14: Temperature anomaly averages (relative to self-climatologies 1961-90) in 20-yr windows for different data sets for different seasonal-regional averages (a = European Arctic in winter, b = Western Siberian Arctic in spring, c = Pacific Arctic in summer, d = eastern Canadian Arctic in autumn). The solid line gives the mean value of all observation based data sets, the horizontal bars (slightly displaced in the vertical for better visualization) indicate the spread. A full version of this figure (including symbols for each data set) is given in the electronic supplement).